### Data Quality

#### Summary

As mentioned above, the established data files (from 2014 to 2019) provided by Dr. Drew processed from raw data\*. Further data wrangling efforts are conducted to (1) make the data files more reproducible for use with machine learning modeling and (2) combine the data files based on calendar years together so we can find trends. We will go into more detail below.

In process (1) we will tackle data wrangling. We refer to Dr. Drew’s data cleaning scripts\*\* including multiple impute and cleaning process. During exploratory data analysis, we manipulated categorical variables to remove outliers. We re-mapped the categorical variable, category\_cd, so it could be reproducible. We also filtered the data on school\_type\_txt which showed an overwhelming favor to Regular Schools. We that, we separated a new data frame just to Regular Schools. The final significant data manipulation is to SPG Score. We filtered out 0’s as it does not make sense to train a model with data that is missing a response variable.

For our process (2) we merged all year-based data files together between 2014 and 2017. Because discrepancies in data 2018/2019, including incomplete data, an inner join process is conducted and the resulted combined dataset that totaled 259 attributes and 9731 rows.

##### References

\*[http://nbviewer.jupyter.org/github/jakemdrew/EducationDataNC/blob/master/2014/School%20Datasets/PublicSchools2014.csv]( http://nbviewer.jupyter.org/github/jakemdrew/EducationDataNC/blob/master/2014/School%20Datasets/PublicSchools2014.csv)

\*\*[https://nbviewer.jupyter.org/github/jakemdrew/EducationDataNC/tree/master/2014/Machine%20Learning%20Datasets/Source%20Code/PublicSchools2014\_ML.ipynb](https://nbviewer.jupyter.org/github/jakemdrew/EducationDataNC/tree/master/2014/Machine%20Learning%20Datasets/Source%20Code/PublicSchools2014\_ML.ipynb)

SPG Grade has 370 records that include "0". These records are only for schools that are listed as Alternative Education or Special Education. The state of North Carolina does not give these school types a score and therefore they were removed from the data set. The predictor "EVAAS Growth Status" has 562 records with "0". The "Alternative Education" school type make nearly half of these records and is the basis for removing these records. Fortunately the data does not have duplicates or any major outliers to address. Once these minor changes have been completed, now the data is tidy and well prepared for EDA.

Our primary data source is a group of separate data files compiled by Dr. Drew for years 2014-2019. For convenience, we have combined the contents of the separate files together into one master file. Unfortunately, because the structure of the data is substantially different for years 2018 and 2019, these records are omitted from the master file.

516 records, or 5% of the data overall, have SPG Scores (our target variable) of 0 (see Visualize Attributes section for more information). Perhaps uncoincidentally, these values all correspond with a small group of elementary schools. While such a finding merits investigation, for purposes of describing factors that influence SPG scores, we will ignore these records for now.

<!-- TODO -->